DOCUMENT RESUME

ED 043 222 EM 008 361

AUTHOR	Bond, Nicholas A., Jr.; Rigney, Joseph W.
TITLE	Measurement of Training Outcomes.
INSTITUTION	University of Southern California, Los Angeles. Dept. of Psychology.
SPONS AGENCY	Office of Naval Research, Washington, D.C. Personnel and Training Branch.
REPORT NO	TR-66
PUB DATE	Jun 70
NOTE	49p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$2.55
DESCRIPTORS	Achievement, *Education, Evaluation, *Measurement, *Research Methodology, Social Change, Training, Training Objectives, Training Techniques

ABSTRACT

Measurement of training outcomes is a requirement for evaluating new training techniques, but is one that is different to meet. Managers of education and training may have different concepts of what they want, as favorable outcomes, than do the investigators doing the research. Classical statistical and experimental designs assume laboratory rigor of control over variables that is seldom possible in the real world of a school or classroom. Yet in the broader perspective of educational institutions, the effectiveness of the institutions is a current issue of fundamental concern in our society. In this report, possibilities for measuring outcomes of training are surveyed, considering training as a form of planned social change. Various approaches are discussed. Illustrations from the computer-assisted instruction (CAI) literature of recent attempts to measure training outcomes are given. The principal conclusions presented are that the classical four-way design is impracticable for most evaluation studies in training environments; that a policy of "adaptive research for big effects" is apt to be scientifically and administratively desirable; and that current attempts at measurement of training outcomes sill use fairly simple methods. (Author)

ED043222

BEHAVIORAL TECHNOLOGY LABORATORIES

Technical Report No. 66

MEASUREMENT OF TRAINING OUTCOMES

June 1970

Department of Psychology
University of Southern California

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

EM008361

ED043222

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

DEPARTMENT OF PSYCHOLOGY

UNIVERSITY OF SOUTHERN CALIFORNIA

Technical Report No. 66

MEASUREMENT OF TRAINING OUTCOMES

June 1970

Nicholas A. Bond, Jr.
Joseph W. Rigney

Prepared for

Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

Contract N00014-67-A-0269-0012
Contract Authority Identification No. NR 154-295

Reproduction in whole or in part is permitted
for any purpose of the United States Government

THIS DOCUMENT HAS BEEN APPROVED FOR PUBLIC
RELEASE AND SALE; ITS DISTRIBUTION IS UNLIMITED

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Behavioral Technology Laboratories
University of Southern California
Los Angeles, California 90007

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE

MEASUREMENT OF TRAINING OUTCOMES

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Technical Report 66 June 1970

5. AUTHOR(S) *(First name, middle initial, last name)*

Nicholas A. Bond, Jr.
Joseph W. Rigney

6. REPORT DATE
June 1970

7a. TOTAL NO. OF PAGES
34

7b. NO. OF REFS
25

8a. CONTRACT OR GRANT NO.
N00014-67-A-0269-0012

b. PROJECT NO.
NR 154-295

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
Technical Report 66

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. DISTRIBUTION STATEMENT
THIS DOCUMENT HAS BEEN APPROVED FOR PUBLIC RELEASE AND SALE;
ITS DISTRIBUTION IS UNLIMITED

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

13. ABSTRACT

Measurement of training outcomes is a requirement for evaluating new training techniques, but is one that is difficult to meet. Managers of education and training may have different concepts of what they want, as favorable outcomes, than do the investigators doing the research. Classical statistical and experimental designs assume laboratory rigor of control over variables that is seldom possible in the real world of a school or classroom. Yet in the broader perspective of educational institutions, the effectiveness of these institutions is a current issue of fundamental concern in our society. In this report, possibilities for measuring outcomes of training are surveyed, considering training as a form of planned social change. Approaches which are discussed include the classic Solomon four-group design, iterative adaptation to the peculiarities of individual student progress, response surface designs, adaptive control models, decision theory models, and simulation models. Illustrations from the CAI literature of recent attempts to measure training outcomes are given. The principal conclusions presented are that the classical four-way design is impracticable for most evaluation studies in training environments; that a policy of "adaptive research for big effects" is apt to be scientifically and administratively desirable; and that current attempts at measurement of training outcomes still use fairly simple methods.

DD FORM 1 NOV 65 1473 (PAGE 1)
S/N 0101-807-6801

Unclassified
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Training Outcomes						
Measurement of Change						
Experimental Designs						
Adaptive Models						
Computer-aided Instruction						
Instructional Technology						

ACKNOWLEDGMENTS

This report is a product of a continuing research program sponsored by the Personnel and Training Research Programs Branch of the Psychological Sciences Divisions, Office of Naval Research. The support, encouragement, and patience of Dr. Victor Fields and Dr. Glenn L. Bryan are gratefully acknowledged.

Portions of this report were given as a lecture at an Advanced Study Institute, sponsored by NATO, at the Royal Naval College, Greenwich, England, in April 1970.

ABSTRACT

Measurement of training outcomes is a requirement for evaluating new training techniques, but is one that is difficult to meet. Managers of education and training may have different concepts of what they want, as favorable outcomes, than do the investigators doing the research. Classical statistical and experimental designs assume laboratory rigor of control over variables that is seldom possible in the real world of a school or classroom. Yet in the broader perspective of educational institutions, the effectiveness of these institutions is a current issue of fundamental concern in our society. In this report, possibilities for measuring outcomes of training are surveyed, considering training as a form of planned social change. Approaches which are discussed include the classic Solomon four-group design, iterative adaptation to the peculiarities of individual student progress, response surface designs, adaptive control models, decision theory models, and simulation models. Illustrations from the CAI literature of recent attempts to measure training outcomes are given. The principal conclusions presented are that the classical four-way design is impracticable for most evaluation studies in training environments; that a policy of "adaptive research for big effects" is apt to be scientifically and administratively desirable; and that current attempts at measurement of training outcomes still use fairly simple methods.

TABLE OF CONTENTS

Section		Page
I.	INTRODUCTION	1
II.	SPECIFIC PERFORMANCE MEASURES	3
	Gain Scores	4
	Number Solved vs. Process Scores	7
	Time to Criterion.	7
	Error Rate	8
	Persistence Measures	8
	Transfer Measures.	9
	Time vs. Achievement	12
	Retention Measures	12
	Remarks.	13
III.	COMPARATIVE DESIGNS FOR EVALUATING TRAINING	14
	Response-surface Designs	20
	Adaptive Control Models	21
	Decision Theory Models	22
	Simulation Models.	23
	Remarks.	24
IV.	ILLUSTRATIONS FROM THE CAI LITERATURE	28
	Remarks.	31
	REFERENCES	33

LIST OF TABLES

Table		Page
1.	Main Types of Criteria for Assessing the Suitability of a Program	2
2.	Average Grade-placement Scores on the Stanford Achievement Test: California, 1966-67	28
3.	Average Grade-placement Scores on the Stanford Achievement Test: California, 1967-68	29
4.	Learning and Test Scores for Three Experimental Treatment Groups	31

LIST OF FIGURES

Figure		Page
1.	Classical four-group design for evaluating training effects	15
2.	Action alternatives, likelihoods, and payoffs in simple decision model	22
3.	Portion of a model simulating a department store buyer's stock ordering behavior	25
4.	Kelley and Prosin's adaptive measurement model	26
5.	Student performance for the portion of the fall quarter final examination in first-year Russian that was common to the computer-based and regular sections	30

MEASUREMENT OF TRAINING OUTCOMES

SECTION I. INTRODUCTION

When somebody says "how effective is training program X," he can be asking about several aspects of X. He may want to know whether the X material covers the subject matter domain which is to be taught, whether it does actually teach whatever it is supposed to teach, whether it is a practical program, and so forth. Thus "effectiveness" is apt to be a multi-dimensional concept with many ramifications. Consider one of Lumsdaine's tables (Table 1), which shows the kinds of internal and external criteria that could be used in evaluating teaching programs, (Lumsdaine, 1965).

In this report, we are mainly concerned with those external criteria that Lumsdaine subsumes under his "effectiveness" category --- that is, with those items of information which show how well the teaching objectives are realized in the students receiving the treatment. At a few places we do touch upon "appropriateness" and "practicality" matters. There are three areas for us to consider in this introductory report: (1) the factors involved in deciding upon specific performance criteria, (2) the selection of some comparison design for showing effectiveness, and (3) some examples from the training evaluation literature. We cannot provide here a cookbook to solve local decisions of scoring and design; what we can do is to raise some of the issues that might give a basis for such decisions. It turns out that, although training effectiveness studies have been rather conventional so far, evaluation schemes derived from adaptive control and from decision theory show some promise for the CAI training manager.

Table 1

Main Types of Criteria for Assessing the Suitability of a Program

	INTERNAL INFORMATION (Available from program inspection)	EXTERNAL INFORMATION
Appropriateness:	Content inspection Table of contents Reading of program Analysis of terminal-behavior frames Publisher's statements of objectives and tests provided by publishers	Stated objectives Test content Competence of authors Opinions of reviews or advisors
Effectiveness:		
1. Predictive criteria	Features of program style or construction Inferred direct and side effects based on program inspection	Reviewer's opinions Developmental history, including tryout and revision Error rates within the program
2. Validating criteria	----------	Measured effects of program use and related data (time, etc.)
Practicality:	Ease of using Reusability Machine (instrumentation) requirements	Cost factors: program price, adaptability, characteristics of presentation, machine (if required)

SECTION II. SPECIFIC PERFORMANCE MEASURES

At first glance it appears that there are many logically possible indexes for measuring learning, and that one could assemble a great array of indexes for the same performance. Fortunately, though, only a few criteria seem to have found much practical use, and choices among them can often be made rather easily. The list of learning criteria below, taken from Bunch (1966), is typical:

"1. A high degree of accuracy in the performance of the response learned.

2. A significantly shorter reaction latency than occurred at the beginning of practice.

3. An increase in the rate or speed of the correct response.

4. An increase in the amplitude of the response.

5. Increased resistance to experimental extinction.

6. Increased resistance to retroactive inhibition from subsequent learning as compared to the amount occurring when learning has been stopped short of mastery.

7. Increased positive transfer to subsequent learning in similar situations.

8. A certain degree of generalization to similar stimulus events."

The phrasing of the above eight criteria is perhaps more reminiscent of academic psychology than of technical training, but the list does serve as a point of departure for the training analyst, and most of the indexes used in training are variants of these eight features. We turn now to the issues that arise when general learning measurements, such as those in Bunch's list, are applied to practical training.

Gain Scores

A test is administered before training, and each man receives an initial score. After a period of training, the same (or equivalent) test is given again, and a final score is obtained. Now there is a "difference score" between these two occasions, and there are at least five ways to handle it. This listing comes from Cattell (1966, p. 388).

"1. Take the difference of before-and-after measures in the raw score units of the given scale.
Change = $a_{2i} - a_{1i}$.

2. Take the difference in standard scores, standardized afresh (separately) for before and after.
Change = $z_{2i} - z_{1i}$.

3. Take the difference in standard scores for the common population sample constituted by before and after together.
Change = $z_{2i} - z_{1i}$.

4. Find the regression of the first score on the second, w, and, as in analysis of covariance, subtract the regression estimate of the second score from the second score.
Change = $a_{2i} - wa_{1i}$.

5. Take the mean of the first and second scores as the better estimate of the individual's typical absolute level and measure his change <u>on each</u> occasion from that. "

As might be expected, each of these five transformations has certain advantages and drawbacks. The "simple raw difference" score of (1) is probably most satisfactory when the test employed is a standard one with pretty good scale characteristics; there may be some published norming data on expected growth over time which can be used to sharpen the analysis. Crude gains, though, are often correlated

negatively with initial score. And when this occurs there is the possibility that a curvilinear relation exists between "amount of knowledge" and measured test score. Lord (1958) says:

> "... the gains of the good students do tend to be numerically less than those of the poor students. However, who is to say but that a gain from an initial true score of 65 to a final true score of 70 may not in every important sense be "greater" than the numerically larger gain from 45 to 55? The former gain, for example, may represent more hours of study or more effort on the part of the teacher or perhaps a more important insight than the latter, numerically larger, gain."

Carver (1966) has explored certain aspects of the relation between amount of knowledge (what is learned) and final exam (what is measured). At least, such explorations should occasion some reservations about acceptance of simple crude gain scores.

Alternative (3) can demonstrate whether or not a general shift occurred over the two testings, and is recommended by Cattell when normalized individual scores are used.

> "Our main practice has been to plot the before-and-after measures ... in a single distribution, which is then normalized. The difference scores are then calculated from absolute scores already thus normalized in this sample from a broader population. By such parameters of the results - agreement of patterns from different experiments, goodness of simple structure, etc. - as running observation has offered, this yields a better approach than any other to good scaling of difference scores." (Cattell, 1966, p. 368)

Alternative (4), first proposed by Manning and Dubois (1962) calculates a "residual gain" score by subtracting from the final score that portion which is predicted by the regression of the first score

on the second. This means that residual gain will correlate zero with initial score. Residual gain does appear to be more predictable from earlier measurements than the simple difference score, and it is therefore of great interest to a technical training agency. Thus (4) should be a "natural" for computer-managed instruction (CMI), though to our knowledge, it has not yet been used in a practical CMI setting. Whatever gains transformation is used, the trainer in describing his results should preserve the raw initial and final scores for each subject so that alternative analyses can be attempted.

Gains or difference scores are apt to be less reliable than either of the scores themselves; this is reflected in the reliability formulas, and leads to one of the "dilemmas" of change measurement: as the correlation between initial and final score increases, reliability of the difference score will decrease. But then in order to have an algebraically reliable difference score, the correlation between initial and final testing would have to be nearly zero, which might indicate that you are not measuring the same thing on the two occasions. So you can have a statistically reliable but meaningless score, or an unreliable, relatively meaningful one (Webster and Bereiter, 1963). Perhaps the dilemma cannot be escaped; but one can, as Cattell suggests, plan to make tests long enough to increase the reliability in the separate scores. Other <u>experimental</u> means of increasing reliability should be adopted where it is possible to do so, (Cattell, 1966, p. 370). In CAI, there is a clear conflict between the requirement for lengthy tests in order to get reliable scores, and the need for <u>short</u> tests in order to keep the pupil from being interminably tested. So far, only the crudest guidelines are available for effecting a tradeoff

between these demands. Indeed, a CMI program could consider this as one of its analytical tasks to be investigated.

Number Solved vs. Process Scores

For long problems which involve many operations in a chain, it is possible to compute process scores as well as overall success scores. The correlations are apt to be moderately high but not perfect. An illustration comes from one of the detailed studies of troubleshooting actions in a simulated electronics environment; if each student "test" (e.g., voltage or resistance) is scored according to its "information value" in reducing the number of alternatives, then students who make the "most informative" checks do tend to get more problems. The general experience, though, is that an overall "number solved" score is most often used, perhaps because it is the easiest to record. The simple success-fail notion is also readily communicable to management, which really "wants a job done within a reasonable time," and is not directly concerned with the elegance of the solution.

Time to Criterion

Since many CAI and CMI programs provide for branching and individuation of response, times to completion, or times to some criterion such as "eight out of ten problems correct," may differ widely among the students in a class. Good practice in handling time scores usually includes some kind of graphic tabulation in addition to the ordinary descriptive statistics, since skewed and truncated time score distributions are frequently observed. Time scores are often very susceptible, too, to short-term motivational factors, so such scores might be

indicated if a training manager was trying out some incentive scheme. Several programs to teach a second language are in use or in development, and time-to-complete distributions for such programs should be important indicators of program effectiveness, assuming that the material covered is comparable to, say, a semester or year course.

Error Rate

Programmed learning practitioners seem to agree that a low error count is a necessary but insufficient condition for learning, (Lumsdaine, 1965). The prompted-frame error rate can be made as low as desired, but if it is fixed so that few errors are made by the slowest students, it hardly can be optimal for the faster learners. A case can be made for tabulating errors throughout a series of learning attempts:

> "As measures of learning there is good reason for regarding success-or-error scores as superior to latency of response, rate, or amplitude of response, in view of the fact that, first, the occurrence of errors in the initial performance of the act constitutes the best evidence that the subject is confronted with a problem for which he does not already have a ready-made response that is correct, and second, the correct performance without error after training provides the best evidence, or behavior measure, that the problem has been mastered. It is also true, generally, that as errors are eliminated during practice, later trials are completed in less time than was required in the early trials. However, if time scores are the only scores available in a learning experiment, interpretation is difficult and questionable." (Bunch, 1966)

Persistence Measures

The fact that a large fraction of students will finish a training sequence, without abnormal prodding, may itself be offered as an indication of program effectiveness. Strangely enough, very few data

have been published which show how likely a student is to finish a program textbook or CAI course on his own. (For a couple of years, the writer urged all students in his advanced statistics course to complete a programmed textbook during the first week of class. This practice, it was hoped, would furnish a quick review and would bring all the students "up to the same level." The recommendation was dropped when it was discovered only one out of twenty or thirty students did more than a few pages of the programmed text. And this was a well-written, well-edited program put out by a major publisher).

Transfer Measures

A program writer may hope that his instruction will not only be effective in his particular teaching situation, but will generalize to other situations as well. Such expectations have sometimes been held for courses in trouble shooting logic; thus, Schuster (1963) gives a general "bracketing" method for isolating troubles. The logic is general and, once mastered, ought to be widely applicable. The present writer was involved in a pilot study wherein a trial subject quickly learned to perform fault localization in a transceiver *via* a computer terminal and a special maintenance logic diagram. Since the approach "worked" for that special situation, and the subject was so enthusiastic about it, we expect that the student might indeed "try to do the same thing" with other equipments if he had the same kind of supporting materials (Rigney, *et al.*, 1966). But the data requirements for proving that transfer occurs are quite severe (some of the experimental issues are discussed in the next section of this report), and so we have little data that are convincing. A claim for transfer effects should be

accompanied by evidence that competing explanations are less likely than the alleged transfer. When well-controlled transfer studies are attempted, they are often negative; a host of studies shows that school or college achievement, for example, is not very predictive of outside-school achievement. Such results encourage us to enunciate a couple of rules of thumb: (1) positive transfer is often much smaller and less reliable than trainers imagine it is; and (2) transfer of "logic" or "system" or "theory" across situations is apt to be facilitated by staged or dimensionalized practice that "moves toward" the desired situation. To illustrate both these rules, we can consider a typical electronics technician school. The electronics taught to a class of technicians will <u>not</u>, in all probability, result in acceptable corrective maintenance performance of new graduates who go into the field. This is probably due to the fact that trainees actually do little trouble shooting in the school; the focus will be on theory and "understanding" rather than on search practice. If search <u>is</u> explicitly taught, the transfer will improve somewhat; if the sample of troubles employed in training is typical of those encountered in the field, even more transfer should occur; and if enough search practice is given for the technicians to attain real fluency down to, say, a circuit-board level, then graduates may in fact be pretty good field troubleshooters when they enter the field.

Gagne (1961) and his associates started from their "hierarchy of learning sets" idea and laid out a basic ordered structure of tasks, proceeding from lower subordinate sets to higher ones, and on up to the ultimate end behavior. This structure leads to a theory of transfer from one set to another.

> "Four logical patterns exist: passing both the higher learning set and the supporting lower set (+ +); failing both the higher set and the lower supporting set (- -); passing the higher and failing the lower (+ -); failing the higher and passing the lower (- +). Gagne's theory predicts higher positive transfer from a recalled learning set and attainment of the adjacent higher relevant learning set. Obviously, either passing both the higher and lower set or failing both would be in accord with the theory (+ + and - - patterns). Passing a higher set after having failed a related subordinate set is directly opposed to the theory (+ - pattern). Failing a higher set after passing a lower set (- + pattern) is not in opposition to the theory, but is taken by Gagne as being partially due to inadequacies in the instructional program. A measure of the proportion of positive transfer may be obtained by summing the (+ +), (- -), and (+ -) pattern and dividing this sum into the sum of the two patterns in accord with the theory (+ +; - -). Ratios approaching 1.00 provide strong confirmation of the theory."
> (Evans, 1965, p. 409).

When subject matter can be arranged into such a structure, transfer predictions ought to be quite reliable. The empirical verification of transfer structure via CMI may be realized in several technical training areas over the next decade.

Mayo (1966) proposes a specific transfer criterion for theory courses in advanced equipments: a theory course is good if it permits rapid learning of operational equipment. This criterion arose because of the practical effects of increased equipment complexity. A recruit could no longer receive general theoretical training in his occupational area and then learn specific equipments on the job --- he had to learn theory on a definite equipment before he could even begin to work on it. The criterion has already been utilized in some U.S. Navy avionics courses.

Time vs. Achievement

Let us say that CAI teaching program X takes longer to complete than competing program Y, but it also results in more student learning than Y. There is then an interpretive problem of putting achievement and time in the same effectiveness formulation. Operations researchers would, perhaps, favor achievement/minute efficiency ratios, but such indexes might jump around because of the achievement scales features (Lumsdaine, 1965, p. 310). Lumsdaine suggests the following as state-of-the-art:

> "1. Report gains in attainment of outcome achieved by going through the program from beginning to end and separately report time spent on the program as a second, separate dependent variable.
>
> 2. Determine and report as the main dependent variable time required to achieve specified levels of attainment.
>
> 3. Hold time constant, reporting attainment achieved in some arbitrarily fixed period of time.
>
> 4. Let both time and attainment vary, using some devised single measure such as amount of attainment per unit time."

Retention Measures

A few training researchers have followed up technicians some months or years after schooling, and checked on how much technical material they have forgotten. Conventional test scores are most often used, perhaps the same final exam that the men took earlier will be given to them. A typical result is that the men suffer a gradual deterioration in remembering tested material; after two or three years,

their scores will be about half of what they made when they graduated from school. There are exceptions, to this, however; DC electronics theory is remembered pretty well, as studies in both British and America show (Wickens, et al., 1952; Dale, 1967).

Remarks

Our quick survey of specific performance indexes shows that the problem of choosing a performance measure is not so simple as some training people believe, and yet it is not hopelessly complex either. Perhaps the CAI planner would not be far from the mark if he would routinely collect entering and final performance data on several kinds of scores, along with student aptitude and other such information as could be economically collated. If the training context is one in which research into the learning process is an important function of the agency doing the training, then detailed process scores might be worth gathering as well.

In one sense, the choice of an index is simple: we want that index which best predicts the final performance. And since best prediction often comes from the combination of several indicators, we might let the computer choose, by weighting, our indexes for us. Carver (1966) raises this possibility, and the CMI possibilities for storing and multivariate weighting are certainly intriguing. Those scores which do not predict final performance would gradually be dropped by the CMI model.

SECTION III. COMPARATIVE DESIGNS FOR EVALUATING TRAINING

Most training authorities are not satisfied with using performance scores for their own internal operations; eventually there may be the need to demonstrate that X is better than Y. When he gathers data to make such claims, the training manager can be considered to be attempting a special kind of planned change. Indeed, most visions of society demand planned change. It is natural, then, to look first at general social evaluation methods in our approach to training evaluation. To take a specific case, the concept of the "post-industrial society" is one possibility that looms before us. Various writers, such as Daniel Bell, foresee several dimensions of such a society: (1) a service economy; (2) a pre-eminent technical class; (3) the centrality of theoretical knowledge as a prime mover; (4) self-sustaining technological growth; (5) creation of an intellectual technology. Maybe the post-industrial society is already here, and maybe changes on the five dimensions above might be accelerated <u>via</u> appropriate intervention. At a more down-to-earth change level, we might attempt to improve the occupational prospects of urban youth, perhaps by work programs, by bonding "nonbondables," or by subsidizing employers.

Evaluation of any planned change effort is, in theory, a sharply articulated enterprise. The basic methodology is straightforward; the following quote from Belasco and Trice (1967) puts well the main requirements:

"1. A clear statement of the expected results of the change experience. The statement should be in observable terms, including the time span over which a specific result can be measurable.

2. The development of relevant, reliable yardsticks which measure progress toward the stated objectives.

3. Application of the yardsticks in terms of the time span implied by the objective.

4. The establishment of an evaluation design which enables the researcher to distinguish the effects of change from those of other intervening contaminants.

5. The establishment of the kinds and sources of information required to evaluate the change experience in terms of the objective. At least two sources of information should be utilized to minimize bias.

6. A specification and examination of those underlying personality and situational factors which explain the identified change."

Except for number 6, these requirements are simple but severe; of the research schemes proposed to implement them, perhaps the most famous is the Solomon four-group design, where pretests and posttests are given according to the following plan:

	Experimental Group	Control Groups		
	A	B	C	D
Pretest	Yes	No	Yes	No
Treatment	Yes	Yes	No	No
Posttest	Yes	Yes	Yes	Yes

Fig. 1. Classical four-group design for evaluating training effects.

In our training context, "treatment" means "training," of course. Now if subjects are randomly assigned to the four groups, one can distinguish a treatment effect and three potential "contaminant" effects:

Treatment effect:	Compare Posttest B with Posttest D
Test effect:	Compare Pretest C with Posttest C
Passage of time effect:	Compare Posttest D with Pretest A Compare Posttest D with Pretest C Compare Posttest D with average of Pretest A and Pretest C
Interaction effect:	Compare Posttest A with Posttest B Compare Posttest A with Posttest D

This classic four-way design does facilitate inference, and when it is accompanied by appropriate statistical tests it is often recommended to social science and education students as <u>the</u> way to conduct an evaluation of a one-level treatment. Unfortunately, this design, and other variants of it, have difficulties. Among the logical ones is the assumption that simple passage of time and the treatment experience are independent in their effects on the final outcome. It is simply inevitable that interaction between these two factors will take place, thus destroying much of the meaning in the Posttest D comparisons with Pretest on A and/or C. Other problems emerged, too, when Belasco and Trice made a serious attempt to apply the four-way plan to an executive training project. The random assignment process guarantees, one assumes, that the starting points for all groups are the same; yet in the actual study it turned out that A and C scores were significantly different on the pretest (there were about 30 subjects in each group). If this is so, then maybe B and D are also originally different, and the entire comparison system collapses, except for the rather uninteresting test effect on Group C.

Another thing is that the four-way plan demands large numbers of persons if satisfactory samples are to be in each condition. For some projects, this would not be a serious problem, but it does mean that except for well-supported situations there will be small samples to contend with.

A fundamental feature of all classical comparison designs is the strict similarity of treatment from one group to another; in their report, Belasco and Trice say that this requirement was quite demanding, both administratively and emotionally.

Still another difficulty, which is encountered in almost any strict evaluation design, involves the random assignment of people to conditions. Here is what happened to a youth-work program which tried to carry out a controlled research plan (Belasco and Trice, 1967).

> "In essence the program required random assignment of trainees to the various program components ... With major decisions affecting a trainee's assignment delegated to a table of random numbers, the counselors came to believe their training and skills had become superfluous ... Some of the counselors reacted by improvising ways of minimizing the impact of random assignment on some trainees ... Many requests for exceptions were made by counselors, and most were accepted by the study director ... In the fall of 1966 the agency and the Federal department ... agreed that the controls of the research design could be loosened in an informal basis."

Participants in change programs can obscure evaluations in other ways, too. An intriguing case comes from Suppes at Stanford, who was testing a computer-aided instruction (CAI) routine in elementary mathematics. Part of his experimental plan was to provide 8 minutes per day of arithmetic practice <u>via</u> a computer-driven console. Every student at the experimental school had this experience, so the design

was basically a two-group A-C comparison. The control school, however, soon initiated a program (counter-program?) of 25 minutes per day of manual drill on the same kinds of problems. The control school "won," in the sense that their final posttest scores were higher than those of the experimentals -- but as Suppes observes, this means only that 25 minutes of manual drill is better than 8 minutes of CAI drill (Suppes & Morningstar, 1969).

Now some of these difficulties can be met or managed in various ways. One might eliminate the pretest entirely, and thus have a two-group (B and D) comparison at the end. This would increase the precision of the remaining comparisons, and the larger numbers of subjects in the two groups should increase the likelihood of equivalent starting points. Stratification should, in many instances, improve the match and increase the sensitivity even further. There seems to be no easy resolution of the random assignment and standard treatment issues, though randomness can occasionally be achieved through deception of some kind, say by announcing the assignment is on some "publicly acceptable" basis, when it is actually random.

Our conclusion regarding the traditional tight experimental designs, then, is rather pessimistic; perhaps such designs are suited for only the most rigidly-controlled situations, such as brief, intensive, and highly standardized technical training. In real world settings, where the change agent is dealing with complex or threatening variables and objectives, the interactions and practical control difficulties do not seem to be worth the effort.

What are the alternatives? There are some hints from the evaluation studies that have been attempted. Alcoholics treated by a "therapeutic community milieu" were not, in general, affected positively

by the treatment (Belasco and Trice, 1967). However, some 20% (of the patient total of 378) did show improvement, and these "succeeders" seemed to be made up of three classes: "improvers," "maintainers," and "AA joiners." Most importantly, each of the three groups tended to have different personality and demographic traits. Knowing and utilizing such trait information should improve the likelihood of treatment success.

Something similar occurred in a supervisor training study. Again, when evaluated by traditional design comparisons, training effects were slight. But some supervisors did change favorably as a result of the training; some even changed on the basis of the pretest alone. And those who changed favorably after training exhibited a different questionnaire-trait pattern from those who were affected by testing only. The lesson is plain: response to planned change is individualized.

Another idea that may be an administrative advance over the traditional designs is to implement a comparative plan, where several treatments are available, and everybody gets some treatment. The comparative design is something of an answer to the ethical problem of giving some subjects no treatment at all, and there are analytical possibilities for exploring effectiveness over a great range of therapeutic procedures. Properly conceived, the "shotgun" research design can be a good thing.

Cooley and Glaser (1969) are doing this in Pittsburgh with elementary school children; what they call Computer-Managed Instruction is essentially a process of iterative adaptation to the individual peculiarities of student progress.

Planned change projects, then, are looking for more flexibility and for recognition of individual variability in response. What other characteristics should the evaluation model have? Baker (1967) proposes an interdisciplinary approach, with many small tight experiments serving to guide the choices of the next treatment; to the technical training man, this might suggest that subsidiary variables (incentives, etc.) could be worked over. Campbell (1967) believes that the achievement of massive treatment effects is the big thing; you can always analyze a big effect after you have it, and perhaps exploit it further.

Recent engineering and statistical models also exhibit qualities of flexibility and adaptability to changed inputs and processes. We will now scan a few of them to see what suggestions they offer for tracking change.

Response-surface Designs

The "evolutionary" statistical design proposed by Box (1954) does more than just test for effects: it tries to locate a point of optimum yield for the variables under consideration, and prescribes new levels on the basis of early results. The method has been applied in the continuous-process industries, and seems to permit genuine savings, on the order of 10 or 15%. Dimensional requirements are stringent, however, and the variables must be quantitatively scaled so that the response surface can be defined and explored for maxima. These necessities cause McLean, a statistician, to doubt that the method could be used right now in people-project evaluation (McLean, 1967, p. 232).

Adaptive Control Models

You have a complex process you want to run; you do not understand the process well enough to prescribe optional input settings in advance, but you do have an output measure and also quantitative indexes on each of the inputs. Suppose the system starts to run. To improve the output, you can randomly cause variation in each of the inputs, observe the effects on the output, and keep adjusting the inputs *via* feedback loops. Eventually, the input variables that cause pronounced effects will receive the most weight, ineffectual factors will be essentially "weighted out," and the system will tend toward maximization of output. A few adaptive configurations can be said to "learn," in the sense that the weighting can be done so as to set up more than one input *pattern*, and then to discriminate among unknown signals *via* the learned weights.

In engineering applications, the parameters are usually well-defined electrical quantities and there is no measurement problem. But the notion of *randomly* varying the *amount* of treatment is an interesting one, seldom attempted with people variables. There is no logical reason why computer-managed instruction could not *direct* such random variation and adaptive weighting ... provided that the variables are well enough defined to permit program control. Already we can think of several candidates: amount of drill on different sectors of the material to be learned, level of competence achieved before "graduating" to the next teaching level, and rate-of-response indicators. Already there may be enough experience with some of these to encourage immediate CMI application. Pask's "error register" scheme, where practice material is selected by the computer program according to accumulated error counts, has been running successfully for some years now. And Pask has

shown that the adaptive approach results in notably faster learning that does ordinary practice (Lewis & Pask, 1965).

Decision Theory Models

It is often possible to simplify a decision problem by listing all the available action alternatives, assigning to each alternative a payoff number, and then choosing the action with the highest payoff. If payoffs are not known exactly, then they are estimated by multiplying the probability of each state by its payoff. In the simple diagram below, you are free to take either path P_0 or path P_1, and then to choose either A_1 or A_2 at the square boxes. The final states, O_1 and O_2, are indicated with their respective likelihoods and payoffs.

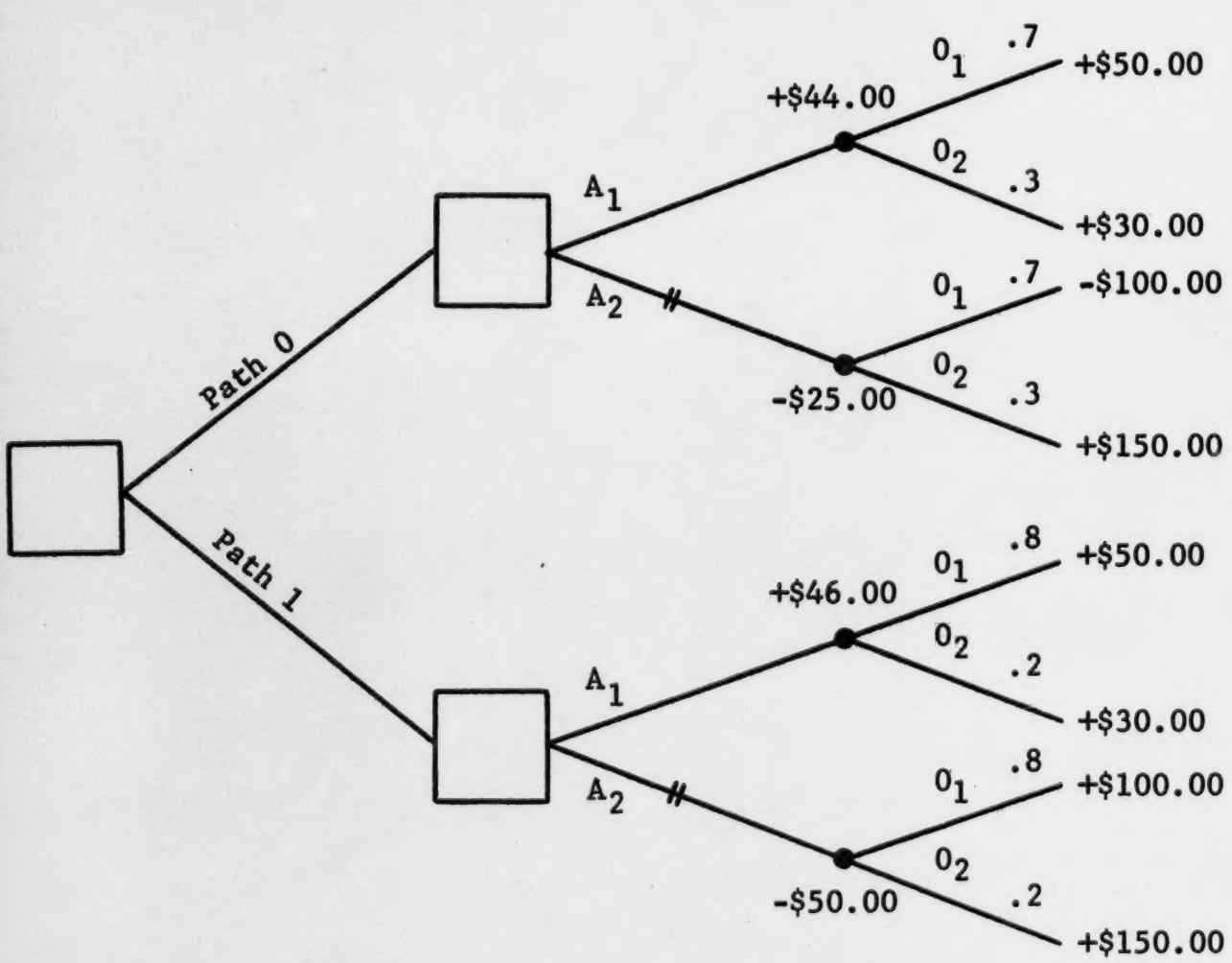


Fig. 2. Action alternatives, likelihoods, and payoffs in simple decision model.

If you take Path 0, your expected return will be $44.00, since your first action on that path will be A_1 and your expected receipts will be (.7) ($50.00) + (.3) ($30.00) = $44.00. Path 1 yields $46.00, so that path would be preferred on a maximum-expected return policy. A_1 and A_2 could in principle be any actions; A_1 might be "provide cash payment of $55.00 a week to trainee without a weekly progress check," whereas A_2 would be "provide $70 a week to trainee if he passes progress check for that week, otherwise pay him nothing." It would cost you, as a training authority, $55.00 and $70.00 respectively to play this game, and whether you should play at all would depend on the end-state probabilities and payoffs. Once those were available, choice could be routinized.

The fact that probabilities of achievement are imperfectly known, or that utilities are only crudely estimated, should not obscure the potentiality of the decision theory model for training management. Since the model is so "clean," it may inspire management to do something about the two central parameters.

Simulation Models

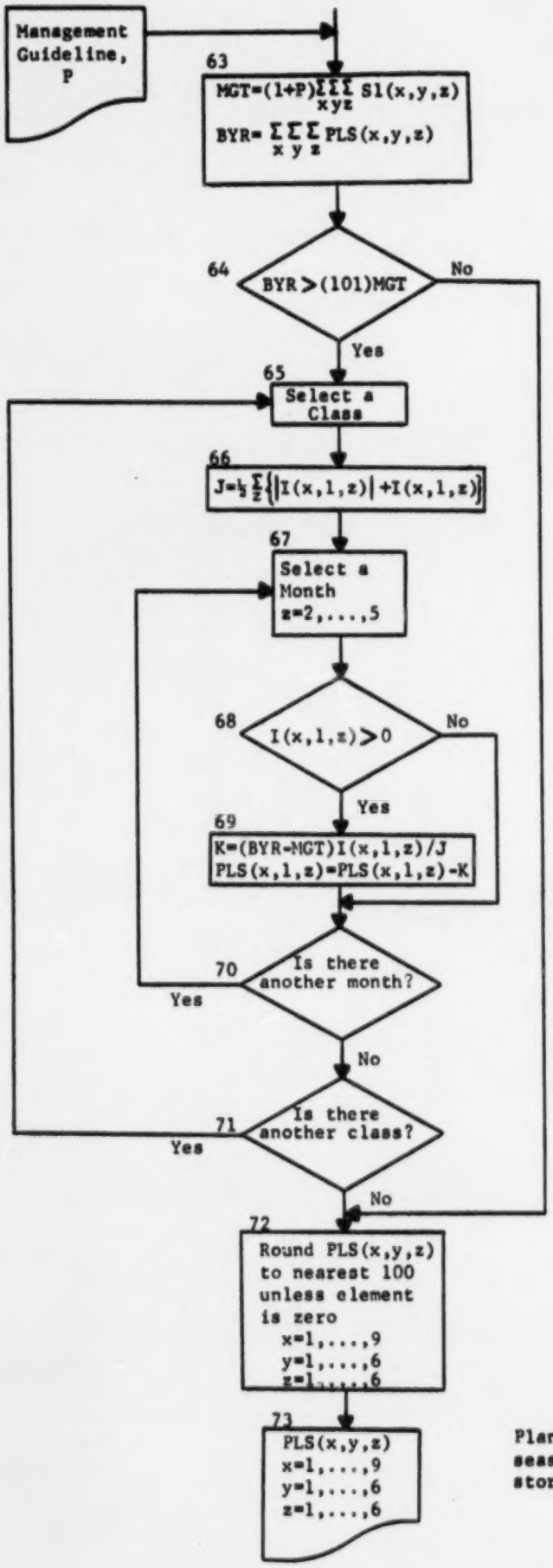
A physical, algebraic, or other representation of a process may be called a simulation if it can imitate some of the behavior of the process. Digital simulation stores information on the time required to complete each step in a process, the likelihood of successful completion, the effects of one task on another, and so on. To accomplish a simulated run through the whole process, sampling of the stored information is performed according to some random-number plan, and the overall performance data are combined. A run ends when the task is completed

or when the allowed time runs out. By changing some of the stored distributions, many questions related to performance can be explored. The technique allows you to "push the limits" of a configuration without the dangers and expense of the real process. Among those processes successfully simulated are aircraft landing behaviors by human pilots, the estuarine flow of the Delaware River basin, and the riding of a bicycle by a computer-driven machine. Figure 3 illustrates a small part of a simulation model of a department store buyer's behavior in ordering his spring ready-to-wear; and comparison of this simulated buyer's decisions with those of real buyers indicates a very good imitation in certain respects.

If we only knew the structural relations that determine complex behaviors we could simulate them. Social scientists such as Bell and Lasswell foresee simulation models of whole societies. It has been seriously suggested that a "Social Planetarium" would improve democracy because citizens could go to it on Sunday afternoon, insert social changes into it, and "see" what the consequences would be. The fact that we are now short of the information needed to do this practically should not dismay us; every theory is a simulation, too, and we are gradually getting better able to measure the variables, and with measurement comes correction and extension of what we already have. Our quick verdict about useful training simulation is: it is closer than many training people imagine, but still far enough away to preclude immediate usefulness.

Remarks

Even this brief glimpse shows us that there are some common and perhaps convergent ideas from these fields of adaptive control, decision



Buyer compares his initial estimates of planned sales with management's sales objective.

Buyer revises his planned sales if they exceed management's objective by 1%.

Sum J of all positive elemental changes is found.

Adjustments are made only if there was an increase in planned sales from sales for last season.

Initial estimate of planned sales is revised for class by month in downtown store to bring his estimate into line. The reduction K in planned elemental sales is computed, and planned elemental sales are revised downward.

Planned sales for spring season by product class, store, and month.

Fig. 3. Portion of a model simulating a department store buyer's stock ordering behavior.

theory, and simulation. One thing in all three models is the idea of a varying and contingent policy to guide action. You do what is best at this time given these conditions, and you may do something different when times and conditions change. Thus all three models assume that the measurement problem is solved, or at least solved well enough to permit decisions on the output. Sometimes the basic ideas can be combined. Look at Kelley and Prosin's (1968) picture of an "adaptive measurement model" (Figure 4).

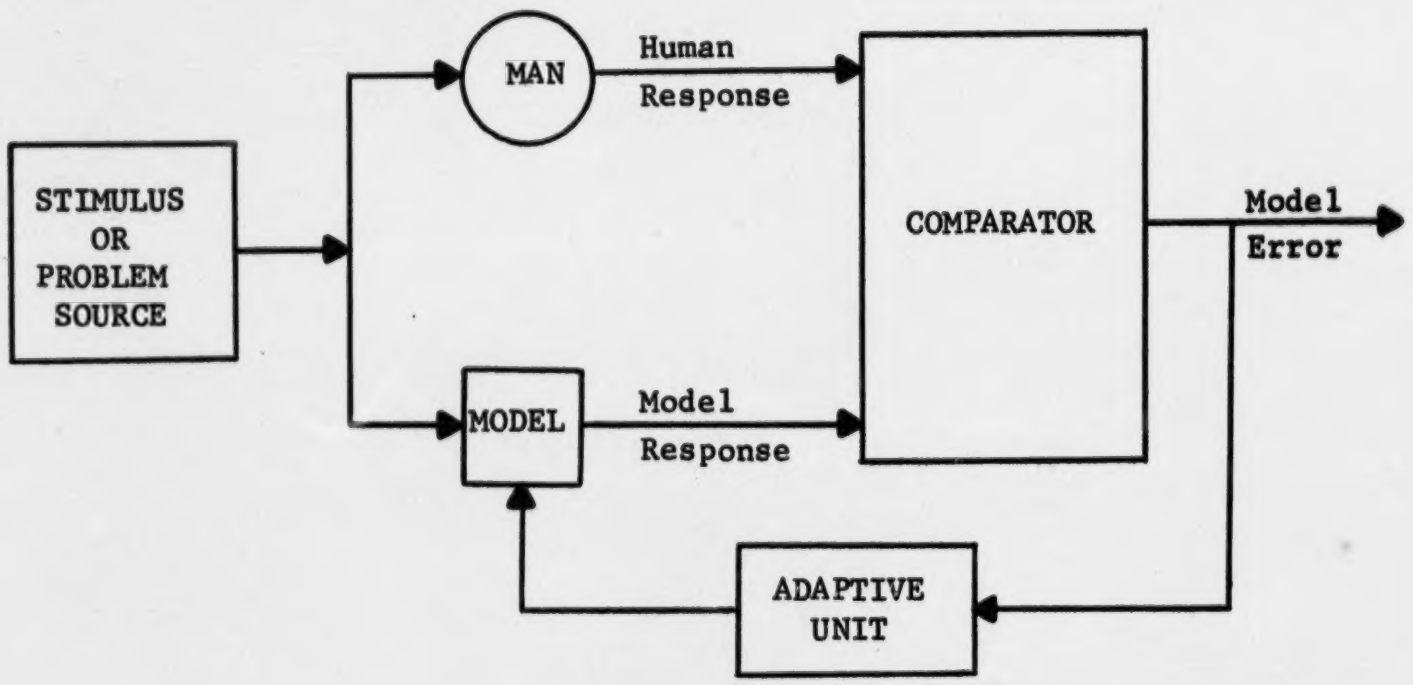


Fig. 4. Kelley and Prosin's adaptive measurement model.

The model adjusts its parameters so it tracks the human operator, and presumably gets to be a better tracker. If we had a training model which was "acceptably good," then it would be a logical next step to insert into that model the cost/payoff features of decision theory. Thus <u>eventually</u> we will be using all three in CAI and CMI.

The decision model requirements for probability and utility numbers do not necessarily preclude fairly immediate application in the CAI context.

Bayesian methods give a fairly good solution to the probability estimation problem. What remains is the utility or scaling problem, and this is tough but not necessarily unmanageable. Raiffa (1968) and others furnish scaling techniques which can at least deal with utility differences across qualitatively different situations. Raiffa, for instance, uses a "standard gamble" choice situation for calibrating utility values for each of three disparate items: a record player, $50.00 and an encyclopedia. Can standard-gamble preferences also yield scale values for the benefits which training people hope will eventuate from their efforts? We believe the possibility is worth a serious trial, if only to determine where the decision approach breaks down.

We conclude with the following stand on designs for training evaluation:

1. The classical four-way design is impracticable for most evaluation studies; a two-group non-pretest design may be an adequate substitute where conditions favor the determination of "one best" treatment.

2. Subjects respond differentially to different treatments; hence multi-factor selection and multi-factor treatments are indicated when possible. Computer-managed instruction can already be of advantage in handling these matters.

3. A policy of "adaptive search for big effects" is apt to be scientifically and administratively desirable. Combination of many (perhaps random) treatment levels, and continuous monitoring of their effects on output would facilitate this search process.

4. A system for estimating outcome probabilities, and a scaling system for calibrating outcome utilities, can provide inputs to a decision model of training choices; this configuration deserves serious trial in practical projects.

SECTION IV. ILLUSTRATIONS FROM THE CAI LITERATURE

Perhaps the most famous CAI project in the late 1960's was the one at Brentwood School in East Palo Alto, California. Several hundred students have now had mathematics instruction in grades 1 through 6 and the program has been extended to other states as well. After two years of trial, Suppes and his associates (1969) published an evaluation of score gains on the Stanford Achievement Test. Here are two tables from the 1966-1967 and 1967-1968 years:

Table 2

Average Grade-placement Scores on the Stanford Achievement Test: California, 1966-67

Grade	Pretest*		Posttest		Posttest-pretest		t	Degrees of freedom
	Experimental	Control	Experimental	Control	Experimental	Control		
			School A versus school B					
3	2.9(51)	3.0(63)	3.9	3.6	1.0	0.6	2.50+	112
4	3.9(60)	3.9(75)	4.7	5.3	0.9	1.4	-2.93+	133
5	4.6(66)	4.6(81)	5.2	6.3	0.7	1.7	-4.74+	145
6	4.9(50)	5.2(70)	7.1	7.1	2.1	1.9	0.95	118
			School C versus school D					
4	3.7(61)	3.8(63)	5.4	4.8	1.7	1.0	4.50+	122
5	5.4(63)	4.9(77)	6.2	5.4	0.8	0.6	1.32	138
6	5.8(58)	6.0(56)	7.4	7.1	1.6	1.1	2.19++	112

*Values in parentheses are numbers of students. +$p < .01$. ++$p < .05$.

Table 3

Average Grade-placement Scores on the Stanford Achievement Test: California, 1967-68

Grade	Pretest*		Posttest		Posttest-pretest		t	Degrees of freedom
	Experimental	Control	Experimental	Control	Experimental	Control		
1	1.39(58)	1.30(267)	2.64	2.51	1.24	1.21	0.33	323
2	2.06(65)	2.16(238)	3.21	2.90	1.15	0.74	5.19+	301
3	3.00(136)	2.85(210)	4.60	3.89	1.59	1.05	6.28+	344
4	3.40(103)	3.49(185)	4.86	5.00	1.46	1.50	-0.38	286
5	4.98(149)	4.44(90)	6.40	5.32	1.42	0.88	4.03+	237
6	5.42(154)	5.70(247)	7.44	7.61	2.02	1.91	0.93	399

*Values in parentheses are numbers of students. +$p < .01$.

Here the evaluative index is crude gain, and the pattern is reasonably clear: experimental (CAI drill) pupils tend to score somewhat better gains. Out of 13 comparisons, 8 are significant in the "right" direction, 2 are "reversed," and the other three are indecisive. A similar table for Mississippi shows an even stronger pattern: six comparisons, six significant gain differences in favor of CAI drill.

The Stanford investigators also give regular CAI training in the Russian language to college students. Each student received 45 or 50 minutes at the console for five days a week. Evaluation in this instance was measured by performance on those parts of the final examination that were common to the regular and the CAI sections. To depict the results, Suppes chose to use errors on the final exam, and also to rank the students in both regular and CAI sections on the final exam. Here is the graph from the fall quarter:

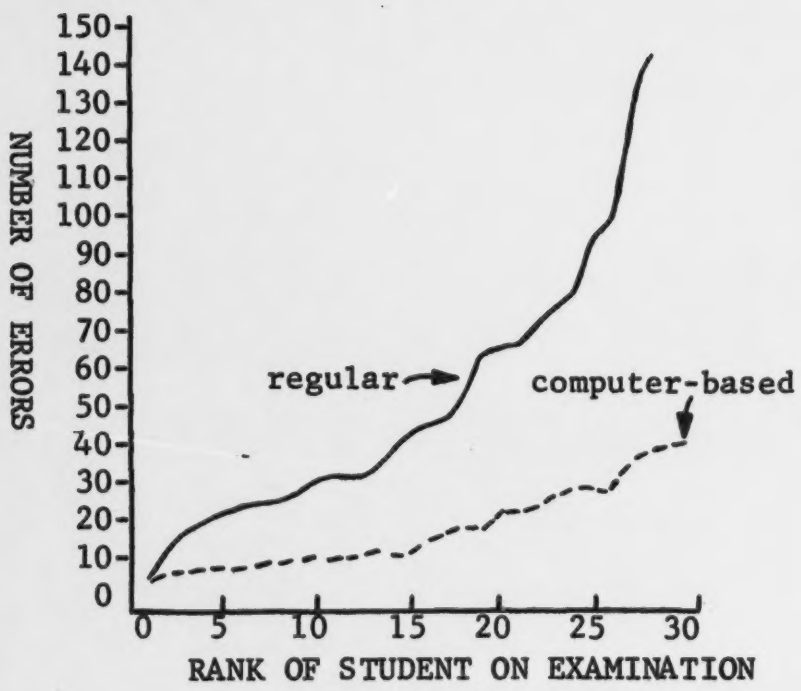


Fig. 5. Student performance for the portion of the fall quarter final examination in first-year Russian that was common to the computer-based and regular sections.

It appears that error scores are much lower for the CAI-instructed students; and that, as rank increases, the CAI superiority becomes more pronounced.

Another measure, and to a language teacher perhaps the most important one, was the percentage of students who finished three quarters of the Russian class under each condition. Twelve of 38 or (22 percent) of the enrollees in the regular class finished, whereas 22 of 30 (73 percent) of the CAI class completed the full three quarters of work. Such differences hardly require statistical tests and we might indeed agree that:

"... this finding suggests that the computer-based course held the interest of the students much better than the regular course did." (Suppes & Morningstar, p. 348).

Franceschi and Hansen thought that CAI might enhance learning via televised instruction. So they gave a lesson in commercial TV program ratings three different ways. Group I received a lecture via instructional television. Group II got the same information via CAI, and Group III viewed the TV tape for a few minutes, went to a CAI terminal for questions on the material just covered. All students took the same posttest, with these outcomes (Hansen & Dick, 1967):

Table 4

Learning and Test Scores for Three Experimental Treatment Groups

	Mean Total Score	Mean Learning Time	Mean Test Time
Group I (TV)	24.31	27:00	21:12
Group II (CAI)	26.56	53:37	19:00
Group III (TV-CAI)	25.75	49:37	22:62

Neither method was much better than the other, it appears, but we do have a case here where the best learning seems to require more time (though a little less test time).

Remarks

These four illustrations of CAI training effects are fairly conservative and do not exhibit any very fancy parameters, tests, or transformations. And yet they are good examples of the evaluations now

being reported. We might expect that, as experience accumulates, more attention will be directed to some of the features we mentioned earlier (residual gains, adaptive scoring, etc.). In his report on the TV-CAI trial, for instance, Hansen gives some correlations between several scoring methods and the total class test (Hansen & Dick, p. 62). This sort of information could be used to improve both the process and end-product scoring procedures.

REFERENCES

Baker, F. B. Experimental design considerations associated with large-scale research projects. In Stanley, J. (Ed.), Improving Design and Statistical Analysis. Chicago: Rand McNally, 1967.

Belasco, J. A., and Trice, H. M. Assessing Change in Training and Therapy. New York: McGraw-Hill, 1967.

Box, G. E. P. The exploration and exploitation of response surfaces: some general considerations and examples. Biometrics, 1954, 10.

Bunch, M. E. The criterion problem in learning research. In Wientge, K. M. & Dubois, P. H. (Eds.), Criteria in Learning Research. St. Louis, Mo.: Department of Psychology, Washington University, 1966 (Tech. Rep. 9, Contract Nonr 816(14)).

Campbell, D. T. Administrative experimentation, Institutional records, and non-reactive measures. In Stanley, J. (Ed.), Improving Experimental Design and Statistical Analysis. Chicago: Rand McNally, 1967.

Carver, R. P. The curvilinear relationship between knowledge and that performance: final examination as the best indicant of learning. In Wientge, K. M. & Dubois, P. H. (Eds.), Criteria in Learning Research. St. Louis, Mo.: Department of Psychology, Washington University, 1966 (Tech. Rep. 9, Contract Nonr 816(14)).

Cattell, R. B. (Ed.) Handbook of Multivariate Experimental Psychology. Chicago: Rand McNally, 1966.

Cooley, W. W. & Glaser, R. The computer and individualized instruction. Science, October 1969, 166.

Dale, H. C. A. The Retention of Basic Electronics Theory. London: Army Personnel Research Committee, APRC 67/ME 1, December 1967.

Evans, J. L. Programming in mathematics and Logic. In Glaser, R. (Ed.), Teaching Machines and Programmed Learning, II. Data and Directions, Washington, D. C.: National Education Association, 1965.

Gagne, R. M. & Paradise, N. E. Abilities and learning sets in knowledge acquisition. Psychol. Monog., 1961, 75(518), 1-23.

Hansen, D. M. & Dick, W. Semi-Annual Progress Report. Tallahassee, Florida: Florida State University CAI Center, July 1967, Report 5.

Kelley, C. R. & Prosin, D. J. Adaptive Performance Measurement. Santa Monica, Calif.: Dunlap & Associates, Final Report Contract Nonr 4986(00), 15 August 1968.

Lewis, B. N. & Pask, G. The theory and practice of adaptive teaching systems. In Glaser, R. (Ed.), Teaching Machines and Programmed Learning, II. Data and Directions, Washington, D.C.: National Education Association, 1965.

Lord, F. M. Further problems in the measurement of growth. Educ. Psychol. Measurement, 1958, 18, 437-454.

Lumsdaine, A. A. Assessing the effectiveness of instructional programs. In Glaser, R. (Ed.), Teaching Machines and Programmed Learning, II. Data and Directions, Washington, D. C.: National Education Association, 1965.

Manning, W. H., & Dubois, P. H. Correctional methods in research on human learning perception and motor skills. Psychol. Rep., Monogr. Suppl. 3-V18, 1962.

Mayo, G. D. Ability to learn operational equipment and systems as a criterion in learning research. In Wientge, K. M. and Dubois, P. H. (Eds.), Criteria in Learning Research. St. Louis, Mo.: Department of Psychology, Washington University, 1966 (Tech. Rep. 9, Contract Nonr 816(14)).

McLean, L. D. Discussion. In Stanley, J. (Ed.), Improving Experimental Design and Statistical Analysis. Chicago: Rand McNally, 1967.

Raiffa, H. Decision Analysis. Menlo Park, Calif.: Addison-Wesley, 1968.

Rigney, J. W., Bond, N. A., Jr., Mason, A. K., & Macaruso, R. B. Training Corrective Maintenance Performance on Electronic Equipment with CAI Terminals. I. A Feasibility Study. Los Angeles: Univer. Southern California, Electronics Personnel Res. Group, December 1966. (Tech. Rep. 51)

Schuster, D. H. Logical Electronic Trouble-Shooting (A Program). New York: McGraw-Hill, 1963.

Suppes, P. & Morningstar, Mona. Computer-assisted instruction, Science, October 1969, 166.

Webster, H. & Bereiter, C. The reliability of changes measured by mental test scores. In Harris, C. W. (Ed.), Problems in Measuring Change. Madison, Wis.: University of Wisconsin Press, 1963.

Wickens, D. D., Stone, G. R. & Highland, R. W. A Study of the Retention of Electronics Fundamentals During Basic Radar Mechanic Training. San Antonio, Texas: Lackland AFB, HRRC Res. Bull. 1952, 52-56.

ONR DISTRIBUTION LIST

4 Chief of Naval Research
Code 458
Department of the Navy
Arlington, Virginia 22217

1 Director
ONR Branch Office
495 Summer Street
Boston, Massachusetts 02210

1 Director
ONR Branch Office
219 South Dearborn Street
Chicago, Illinois 60604

1 Director
ONR Branch Office
1030 East Green Street
Pasadena, California 91101

6 Director
Naval Research Laboratory
Washington, D. C. 20390
ATTN: Library, Code 2029 (ONRL)

1 Office of Naval Research
Area Office
207 West Summer Street
New York, New York 10011

1 Office of Naval Research
Area Office
1076 Mission Street
San Francisco, California 94103

1 Technical Library
U.S. Naval Weapons Laboratory
Dahlgren, Virginia 22448

1 Research Director, Code 06
Research and Evaluation Department
U.S. Naval Examining Center
Building 2711--Green Bay Area
Great Lakes, Illinois 60088
ATTN: C. S. Winiewicz

6 Director
Naval Research Laboratory
Washington, D. C. 20390
ATTN: Technical Information
Division

20 Defense Documentation Center
Cameron Station, Building 5
5010 Duke Street
Alexandria, Virginia 22314

1 Commanding Officer
Service School Command
U.S. Naval Training Center
San Diego, California 92133

3 Commanding Officer
Naval Personnel and Training
Research Laboratory
San Diego, California 92152

1 Commanding Officer
Naval Medical Neuropsychiatric
Research Unit
San Diego, California 92152

1 Commanding Officer
Naval Air Technical Training
Center
Jacksonville, Florida 32213

1 Dr. James J. Regan, Code 55
Naval Training Device Center
Orlando, Florida 32813

1 Chief, Naval Air Technical
Training
Naval Air Station
Memphis, Tennessee 38115

1 Director
Education and Training Sciences
Department
Naval Medical Research Institute
National Naval Medical Center
Building 142
Bethesda, Maryland 20014

1 Chairman, Behavioral Science Dept.
Naval Command and Management Div.
U.S. Naval Academy, Luce Hall
Annapolis, Maryland 21402

1 Chairman, Management Science Dept.
Naval Command and Management Div.
U.S. Naval Academy, Luce Hall
Annapolis, Maryland 21402

1 Dr. A. L. Slafkosky
Scientific Advisor (Code AX)
Commandant of the Marine Corps
Washington, D. C. 20380

1 Behavioral Sciences Department
Naval Medical Research Institute
National Naval Medical Center
Bethesda, Maryland 20014

1 Commanding Officer
Naval Medical Field Research Lab.
Camp Lejeune, North Carolina 28542

1 Director
Aerospace Crew Equipment Department
Naval Air Development Center
Johnsville
Warminster, Pennsylvania 18974

1 Mr. George N. Graine
Naval Ship Systems Command
(SHIPS 03H)
Department of the Navy
Washington, D. C. 20360

1 Chief
Bureau of Medicine and Surgery
Code 513
Washington, D. C. 20390

1 Chief
Bureau of Medicine and Surgery
Research Division (Code 713)
Department of the Navy
Washington, D. C. 20390

1 Commander
Submarine Development Group Two
Fleet Post Office
New York, New York 09501

1 Commander
Operational Test & Evaluation
Force
U.S. Naval Base
Norfolk, Virginia 23511

1 Office of Civilian Manpower
Management
Technical Training Branch (Code 024)
Department of the Navy
Washington, D. C. 20390

1 Chief of Naval Operations (Op-07TL)
Department of the Navy
Washington, D. C. 20350

1 Chief of Naval Material (MAT 031M)
Room 1323, Main Navy Building
Washington, D. C. 20360

1 Library, Code 0212
Naval Postgraduate School
Monterey, California 93940

1 Technical Reference Library
Naval Medical Research Institute
National Naval Medical Center
Bethesda, Maryland 20014

1 Scientific Advisory Team (Code 71)
Staff, COMASWFORLANT
Norfolk, Virginia 23511

1 Education & Training Developments
Staff
Personnel Research & Development Lab
Washington Navy Yard, Building 200
Washington, D. C. 20390

9 Technical Library (Pers-11b)
Bureau of Naval Personnel
Department of the Navy
Washington, D. C. 20370

3 Personnel Research & Development
Laboratory
Washington Navy Yard, Building 200
Washington, D. C. 20390
ATTN: Library, Room 3307

1 Commander, Naval Air Systems Command
Navy Department, AIR-4132
Washington, D. C. 20360

1 Commandant of the Marine Corps
Headquarters, U.S. Marine Corps
Code A01B
Washington, D. C. 20380

1 Technical Library
Naval Ship Systems Command
Main Navy Building, Room 1532
Washington, D. C. 20360

1 Mr. Philip Rochlin, Head
Technical Library Branch
Naval Ordnance Station
Indian Head, Maryland 20640

1 ERIC Clearinghouse on Vocational and Technical Education
The Ohio State University
1900 Kenny Road
Columbus, Ohio 43210
ATTN: Acquisition Specialist

1 LT. COL. F. R. Ratliff
Office of the Assistant Secretary of Defense (M&RU)
The Pentagon, Room 3D960
Washington, D. C. 20301

1 Dr. Ralph R. Canter
Military Manpower Research Coordinator
OASD (M&RA) MR&U
The Pentagon, Room 3D960
Washington, D. C. 20301

1 Dr. Don H. Coombs, Co-Director
ERIC Clearinghouse
Stanford University
Palo Alto, California 94305

1 ERIC Clearinghouse on
Educational Media and Technology
Stanford University
Stanford, California 94305

1 Deputy Director
Office of Civilian Manpower Management
Department of the Navy
Washington, D. C. 20390

1 Chief, Naval Air Reserve Training
Naval Air Station
Box 1
Glenview, Illinois 60026

1 Technical Library
Naval Training Device Center
Orlando, Florida 32813

1 Director
Human Resources Research Organization
300 North Washington Street
Alexandria, Virginia 22314

1 Human Resources Research Organization
Division #1, Systems Operations
300 North Washington Street
Alexandria, Virginia 22314

1 Human Resources Research Organization
Division #3, Recruit Training
Post Office Box 5787
Presidio of Monterey, California 93940
ATTN: Library

1 Human Resources Research Organization
Division #4, Infantry
Post Office Box 2086
Fort Benning, Georgia 31905

1 Human Resources Research Organization
Division #5, Air Defense
Post Office Box 6021
Fort Bliss, Texas 79916

1 Human Resources Research Organization
Division #6, Aviation
Post Office Box 428
Fort Rucker, Alabama 36360

1 Commandant
U.S. Army Adjutant General School
Fort Benjamin Harrison, Indiana 46216
ATTN: ATSAG-EA

1 Director of Research
U.S. Army Armor Human Research Unit
Fort Knox, Kentucky 40121
ATTN: Library

1 Armed Forces Staff College
Norfolk, Virginia 23511
ATTN: Library

1 Director
Behavioral Sciences Laboratory
U.S. Army Research Institute of Environmental Medicine
Natick, Massachusetts 01760

1 U.S. Army Behavior and Systems Research Laboratory
Commonwealth Building, Room 239
1320 Wilson Boulevard
Arlington, Virginia 22209

1 Division of Neuropsychiatry
Walter Reed Army Institute of Research
Walter Reed Army Medical Center
Washington, D. C. 20012

1 Behavioral Sciences Division
Office of Chief of Research and Development
Department of the Army
Washington, D. C. 20310

1 Center for Research in Social Systems
American Institutes for Research
10605 Concord Street
Kensington, Maryland 20795
ATTN: ISB

1 Dr. George S. Harker, Director
Experimental Psychology Division
U.S. Army Medical Research Lab.
Fort Knox, Kentucky 40121

1 Director
Air University Library
Maxwell Air Force Base, Alabama 36112
ATTN: AUL-8110

1 Headquarters, Electronic Systems Division
ATTN: Dr. Sylvia Mayer / ESMDA
L. G. Hanscom Field
Bedford, Massachusetts 01730

1 Commandant
U.S. Air Force School of Aerospace Medicine
ATTN: Aeromedical Library (SMSL-4)
Brooks Air Force Base, Texas 78235

1 AFHRL (TR/Dr. G. A. Eckstrand)
Wright-Patterson Air Force Base
Ohio 45433

1 Personnel Research Division (AFHRL)
Lackland Air Force Base
San Antonio, Texas 78236

1 AFOSR(SRLB)
1400 Wilson Boulevard
Arlington, Virginia 22209

1 Headquarters, U.S. Air Force
Chief, Personnel Research and Analysis Division (AFPDPL)
Washington, D. C. 20330

1 Headquarters, U.S. Air Force
AFPTRBD
Programs Resources and Technology Division
Washington, D. C. 20330

1 AFHRL (HRTT/Dr. Ross L. Morgan)
Wright-Patterson Air Force Base
Ohio 45433

1 Dr. Alvin E. Goins, Executive Secy.
Personality and Cognition Research Review Committee
Behavioral Sciences Research Branch
National Institute of Mental Health
5454 Wisconsin Avenue, Room 10A02
Chevy Chase, Maryland 20015

1 Office of Computer Information
Center for Computer Sciences and Technology
National Bureau of Standards
Washington, D. C. 20234

2 Executive Secretariat
Interagency Committee on Manpower Research
1111 Twentieth Street, N.W., Room 251-A
Washington, D. C. 20036

1 Mr Joseph J. Cowan, Chief
Psychological Research Branch (P-1)
U.S. Coast Guard Headquarters
400 Seventh Street, S.W.
Washington, D. C. 20226

1 Executive Officer
American Psychological Association
1200 Seventeenth Street, N.W.
Washington, D. C. 20036

1 Dr. Bernard M. Bass
University of Rochester
Management Research Center
Rochester, New York 14627

1 Dr. Lee R. Beach
Department of Psychology
University of Washington
Seattle, Washington 98105

1 Dr. Donald L. Bitzer
Computer-Based Education Research Laboratory
University of Illinois
Urbana, Illinois 61801

1 Dr. Lee J. Cronbach
School of Education
Stanford University
Stanford, California 94305

1 Dr. Edward R.F.W. Crossman
Department of Industrial Engineering
University of California
Berkeley, California 94720

1 Dr. Robert Dubin
Graduate School of Administration
University of California
Irvine, California 92650

1 Dr. Philip H. DuBois
Department of Psychology
Washington University
Lindell & Skinner Boulevards
St. Louis, Missouri 63130

1 Dr. Marvin D. Dunnette
University of Minnesota
Department of Psychology
Elliot Hall
Minneapolis, Minnesota 55455

1 S. Fisher, Research Associate
Computer Facility, Graduate Center
City University of New York
33 West 42nd Street
New York, New York 10036

1 Dr. John C. Flanagan
American Institutes for Research
Post Office Box 1113
Palo Alto, California 94302

1 Dr. Robert Glaser
Learning Research and Development Center
University of Pittsburgh
Pittsburgh, Pennsylvania 15213

1 Dr. Albert S. Glickman
American Institutes for Research
8555 Sixteenth Street
Silver Spring, Maryland 20910

1 Dr. Bert Green
Department of Psychology
Johns Hopkins University
Baltimore, Maryland 21218

1 Dr. Duncan N. Hansen
Center for Computer Assisted Instruction
Florida State University
Tallahassee, Florida 32306

1 Dr. M. D. Havron
Human Sciences Research, Inc.
Westgate Industrial Park
7710 Old Springhouse Road
McLean, Virginia 22101

1 Dr. Carl E. Helm
Department of Educational Psychology
Graduate Center
City University of New York
33 West 42nd Street
New York, New York 10036

1 Dr. Lloyd G. Humphreys
Department of Psychology
University of Illinois
Champaign, Illinois 61820

1 Dr. Frederic M. Lord
Educational Testing Service
20 Nassau Street
Princeton, New Jersey 08540

1 Dr. Robert R. Mackie
Human Factors Research, Inc.
Santa Barbara Research Park
6780 Cortona Drive
Goleta, California 93017

1 Dr. Richard Myrick, President
Performance Research, Inc.
919 Eighteenth Street, N.W.,
Suite 425
Washington, D. C. 20036

1 Dr. Stanley M. Nealey
Department of Psychology
Colorado State University
Fort Collins, Colorado 80521

1 Dr. Gabriel D. Ofiesh
Center for Educational Technology
Catholic University
4001 Harewood Road, N.E.
Washington, D. C. 20017

1 Mr. Luigi Petrullo
2431 North Edgewood Street
Arlington, Virginia 22207

1 Dr. Len Rosenbaum
Psychology Department
Montgomery College
Rockville, Maryland 20852

1 Dr. Arthur I. Siegel
Applied Psychological Services
Science Center
404 East Lancaster Avenue
Wayne, Pennsylvania 19087

1 Dr. Paul Slovic
Oregon Research Institute
Post Office Box 3196
Eugene, Oregon 97403

1 Dr. Ledyard R. Tucker
University of Illinois
Psychology Building
Urbana, Illinois 61820

1 Dr. John Annett
Department of Psychology
Hull University
Hull
Yorkshire, England

1 Dr. M. C. Shelesnyak
Interdisciplinary Communications
Program
Smithsonian Institution
1025 Fifteenth Street, N.W.,
Suite 700
Washington, D. C. 20005

1 Educational Testing Service
Division of Psychological Studies
Rosedale Road
Princeton, New Jersey 08540

1 Dr. George E. Rowland
Rowland and Company, Inc.
Post Office Box 61
Haddonfield, New Jersey 08033

1 Dr. Mats Bjorkman
University of Umea
Department of Psychology
Umea 6, Sweden

1 Dr. Victor Fields
Personnel and Training Research
Programs
Office of Naval Research, Code 458
Department of the Navy
Arlington, Virginia 22217

1 Dr. Carl E. Helm
Department of Educational Psychology
Graduate Center
City University of New York
33 West 42nd Street
New York, New York 10036

1 Dr. Albert E. Hickey
Entelek, Incorporated
42 Pleasant Street
Newburyport, Massachusetts 01950

1 Dr. Lloyd G. Humphreys
Department of Psychology
University of Illinois
Champaign, Illinois 61820

1 Dr. Robert R. Mackie
Human Factors Research, Inc.
Santa Barbara Research Park
6780 Cortona Drive
Goleta, California 93017

1 Dr. Richard Myrick, President
Performance Research, Inc.
919 Eighteenth Street, N.W., Suite 425
Washington, D. C. 20036

1 Dr. Stanley M. Nealey
Department of Psychology
Colorado State University
Fort Collins, Colorado 80521

1 Dr. Gabriel D. Ofiesh
Center for Educational Technology
Catholis University
4001 Harewood Road, N.E.
Washington, D. C. 20017

1 Mr. Luigi Petrullo
2431 North Edgewood Street
Arlington, Virginia 22207

1 Dr. Len Rosenbaum
Psychology Department
Montgomery College
Rockville, Maryland 20852

1 Dr. Arthur I. Siegel
Applied Psychological Services
Science Center
404 East Lancaster Avenue
Wayne, Pennsylvania 19087

1 Dr. Paul Slovic
Oregon Research Institute
Post Office Box 3196
Eugene, Oregon 97403

1 Dr. Arthur W. Staats
Department of Psychology
University of Hawaii
Honolulu, Hawaii 96822

1 Dr. Ledyard R. Tucker
University of Illinois
Psychology Building
Urbana, Illinois 61820

1 Dr. Benton J. Underwood
Department of Psychology
Northwestern University
Evanston, Illinois 60201

1 Dr. John Annett
Department of Psychology
Hull University
Hull
Yorkshire, England

1 Dr. M. C. Shelesnyak
Interdisciplinary Communications
Program
Smithsonian Institution
1025 Fifteenth Street, N.W. Suite 700
Washington, D. C. 20005

1 Dr. Harold Gulliksen
Department of Psychology
Princeton University
Princeton, New Jersey 08540

1 Educational Testing Service
Division of Psychological Studies
Rosedale Road
Princeton, New Jersey 08540

1 Dr. George E. Rowland
Rowland and Company, Inc.
Post Office Box 61
Haddonfield, New Jersey 08033

1 Dr. Mats Bjorkman
University of Umea
Department of Psychology
Umea 6, SWEDEN

1 Dr. Howard H. Kendler
Department of Psychology
University of California
Santa Barbara, California 93106

1 Director
Human Resources Research Organization
300 North Washington Street
Alexandria, Virginia 22314